PMut. A web-based tool for the annotation of pathological variants on proteins, 2017 update.

Supplementary Material

Víctor López-Ferrando[1], Andrea Gazzo[2], Xavier de la Cruz[3,4], Modesto Orozco[2,5*], Josep Ll. Gelpí[1,5*]

1. Barcelona Supercomputing Center (BSC). Joint Program BSC-CRG-IRB research Program for Computational Biology. Barcelona. Spain.
2. Institute for Research in Biomedicine (IRB) Barcelona. The Barcelona Institute of Science and Technology. Barcelona. Spain.
3. Vall d'Hebron Institute of Research (VHIR), Universitat Autònoma de Barcelona, Barcelona, Spain.
4. Institut Català de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

* To whom correspondence should be addressed. Tel: 34934034009; Fax: 34934021559; Email: gelpi@ub.edu.

Correspondence may also be addressed to Tel: 34934037156; Fax: 34934037157; Email: modesto.orozco@irbbarcelona.org.

CONTENTS

SUPPLEMENTARY TABLES

Table S1. Features computed by PyMut.
Description of the 215 features computed by PyMut. The sequence conservation features require PSI-BLAST searches and Kalign2 multiple sequence alignments to be computed. These have been precomputed for all human sequences in UniRef100 and can be found in the PMut Web portal in their respective protein page.

Type of features	**Number of features**	**Description**
Substitution matrix score	5	Score of the amino acid substitution in the BLOSUM50, BLOSUM62, BLOSUM80 (1), PAM60 (2), and Miyata (3) matrices.
Physical properties difference	8	Relative and absolute difference in volume (4), hydrophobicity (5) and free energy transfer octanol-water (6); Kyte-Doolittle hydropathy index (7) and position in the protein sequence.
Protein interactome graph topology	6	Descriptors of the protein in the interactome graph: degree and five measures of centrality: betweenness, cross-clique, closeness, eigenvector and degree centrality.
Sequence conservation	196	Sequence conservation features are extracted from 4 different sources: two PSI-BLAST (8) searches over UniRef100 and UniRef90 clusters (9), and two multiple sequence alignments by Kalign2 (10) of the sequences found by PSI-BLAST. Each of these 4 alignments is filtered in 4 different ways: 1) taking all the sequences, 2) keeping only the human sequences, 3) excluding all the human sequences and 4) (only in PSI-BLAST) taking the matches under an stricter e-value threshold. From each of these 14 conservation sources, a set of 14 features is computed, which you can find in the Table S2.

Table S2. Conservation features computed by PyMut.

Number of features	Description
2*	Number of sequences in the alignment.
2*	Number of amino acids in the aligned position (no gaps).
4*	Total and relative number of aligned wild type amino acids.
4*	Total and relative number of aligned mutated amino acids.
2*	Position Weight Matrix score, defined as: $PWM_{wt \to mt} = log\left(\frac{number\ of\ mt}{Freq[mt]}\right) - log\left(\frac{number\ of\ wt}{Freq[wt]}\right)$ **

* Each of these features is computed both in a weighted and unweighted fashion. The weighted features give more importance to the most similar sequences in the alignment. Matches in the PSI-BLAST searches are weighted using the BLAST score and matches in the multiple sequence alignment are weighted using the sequence similarity.

** $Freq[mt]$ and $Freq[wt]$ are the relative presence of the mutated and the wild type amino acids in the complete Swiss-Prot (11) database.

Table S3. Selected features for the PMut2017 predictor.

#	Alignment	Database	Filter	Value	Weighted
1	PSI-BLAST	UniRef100	E-value < 10^{-75}	Number of amino acids in the aligned position.	No
2	PSI-BLAST	UniRef100	E-value < 10^{-75}	Position Weight Matrix score.	No
3	PSI-BLAST	UniRef100	E-value < 10^{-75}	Proportion of wild type amino acids in the aligned position.	BLAST score
4	PSI-BLAST	UniRef90	E-value < 10^{-45}	Number of amino acids in the aligned position.	No
5	MSA (Kalign2)	UniRef100	All	Number of sequences in the alignment.	No
6	MSA (Kalign2)	UniRef100	All	Number of wild type amino acids in the aligned position.	No
7	MSA (Kalign2)	UniRef100	All	Position Weight Matrix score.	Sequence similarity
8	MSA (Kalign2)	UniRef100	Human	Number of amino acids in the aligned position.	No
9	MSA (Kalign2)	UniRef90	All	Number of amino acids in the aligned position.	Sequence similarity
10	MSA (Kalign2)	UniRef90	All	Number of wild type amino acids in the aligned position.	Sequence similarity
11	MSA (Kalign2)	UniRef90	Human	Number of amino acids in the aligned position.	Sequence similarity
12	Miyata substitution matrix score.				

Table S4. Performance comparison of classifiers for PMut2017.
Comparison of the six classifiers included in PyMut using the PMut2017 training set (SwissVar (15) October 2016).

Classifier	Accuracy	Sensitivity	Specificity	AUC	MCC
Random Forest (12)	0.81	0.75	0.86	0.81	0.62
AdaBoost (13)	0.81	0.75	0.86	0.80	0.61
Extremely Randomized Trees (14)	0.80	0.72	0.86	0.79	0.59
Logistic Regression	0.77	0.70	0.83	0.76	0.53
Stochastic Gradient Descent	0.77	0.69	0.82	0.76	0.52
Gaussian Naive Bayes	0.67	0.91	0.49	0.70	0.42

The metrics are the result of a 10-fold cross-validation on protein families with 50% sequence identity exclusion (no sequence in the testing set shares more than 50% sequence identity with any protein in the training set). AUC is the Area under the Receiver Operating Characteristic curve and MCC is the Matthews correlation coefficient.

Random Forest is chosen as the PMut2017 classifier, as it outputs the best predictions, followed by AdaBoost, and is computationally more efficient than the later.

Table S5. Functions and variables of the PyMut module.
List of the most important functions and variables exported by the PyMut module. Other helper functions are provided to perform input/output tasks such as reading SwissVar files, FASTA files, or parsing common mutation formats.

Function or variable	Description
`CLASSIFIERS`	Variable containing the six available classifiers, their name, function and default parameters.
`FEATURES`	List of all 215 features that can be computed by PyMut.
`PMUT_FEATURES`	List of the 12 features selected in the PMut2017 predictor.
`FOLDS`	List of different fold generation strategies for cross-validation: k-fold, stratified k-fold, label-exclusive k-fold, etc.
`compute_features`	Compute features for the given variants.
`features_distribution`	Plot features histograms, separating Neutral and Disease mutations.
`iterative_features_selection`	Select features using the iterative algorithm described in Figure S1.
`cross_validate`	Perform a cross-validation with the provided variants, classifiers and fold generation technique.
`evaluate`	Evaluate a prediction using a standard set of metrics: accuracy, precision, sensitivity, specificity, ROC AUC and MCC.
`roc_curve`	Plot a ROC curve using evaluation data.
`train`	Train a predictor.
`predict`	Predict the pathology of variants using a given predictor.
`get_learning_curve`	Plot learning curve (to estimate how better a predictor can be expected to get by making the training set bigger).

.

Table S6. PyMut software dependencies.
List of dependencies of the PyMut Python 3 module.

Python module	URL	Description
NumPy (v1.10)	numpy.org	Fast numerical computing library.
SciPy (v0.17)	scipy.org	Scientific computing library.
Pandas (v0.17)	pandas.pydata.org	Python data analysis library.
Matplotlib (v1.5)	matplotlib.org	Python plotting library.
Seaborn (v0.8)	seaborn.pydata.org	Statistical data visualization library.
Scikit-learn (v0.17)	scikit-learn.org	Machine learning methods.

These dependencies are documented in the official repository package (https://pypi.python.org/pypi/pymut) and will be installed automatically by the standard Python package manager (pip).

SUPPLEMENTARY FIGURES

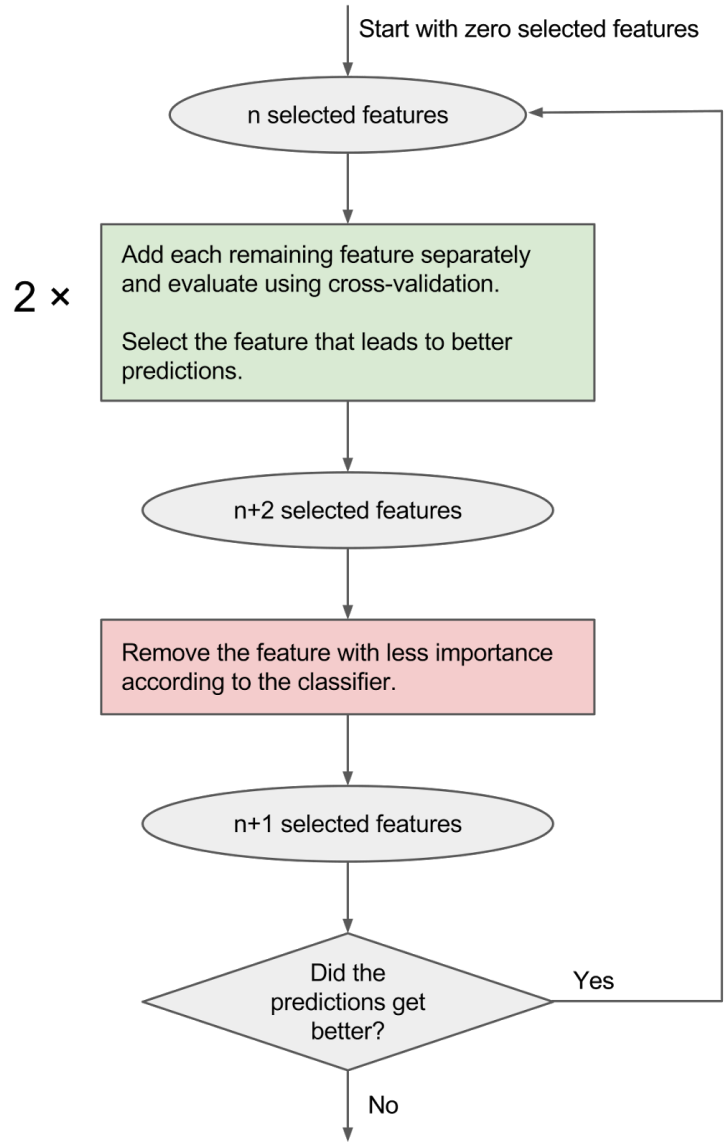

Figure S1. Iterative feature selection algorithm.
Features are added to the selected set until the performance increase in terms of the Matthews correlation coefficient (MCC) is negligible. At each step, the two features that increase the MCC the most are added and then the least important feature is removed. This approach is intended to skip local minima in performance.

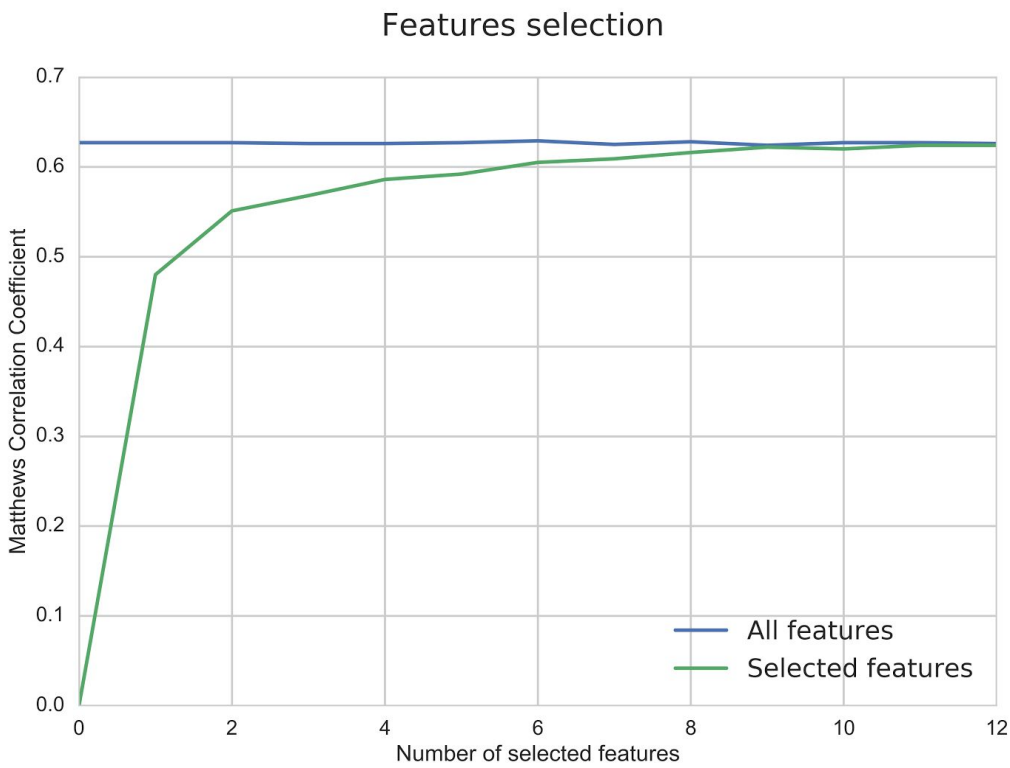

Figure S2. Feature selection for PMut2017.
Feature selection algorithm run for the PMut2017 classifier. The predictor performance increased with each feature added to the selection, and it matched the performance of the predictor using all 215 features with 12 selected features. Note that the variation of the target MCC at each step is due to a change in the cross-validation folds, which are different and randomly chosen at each step.

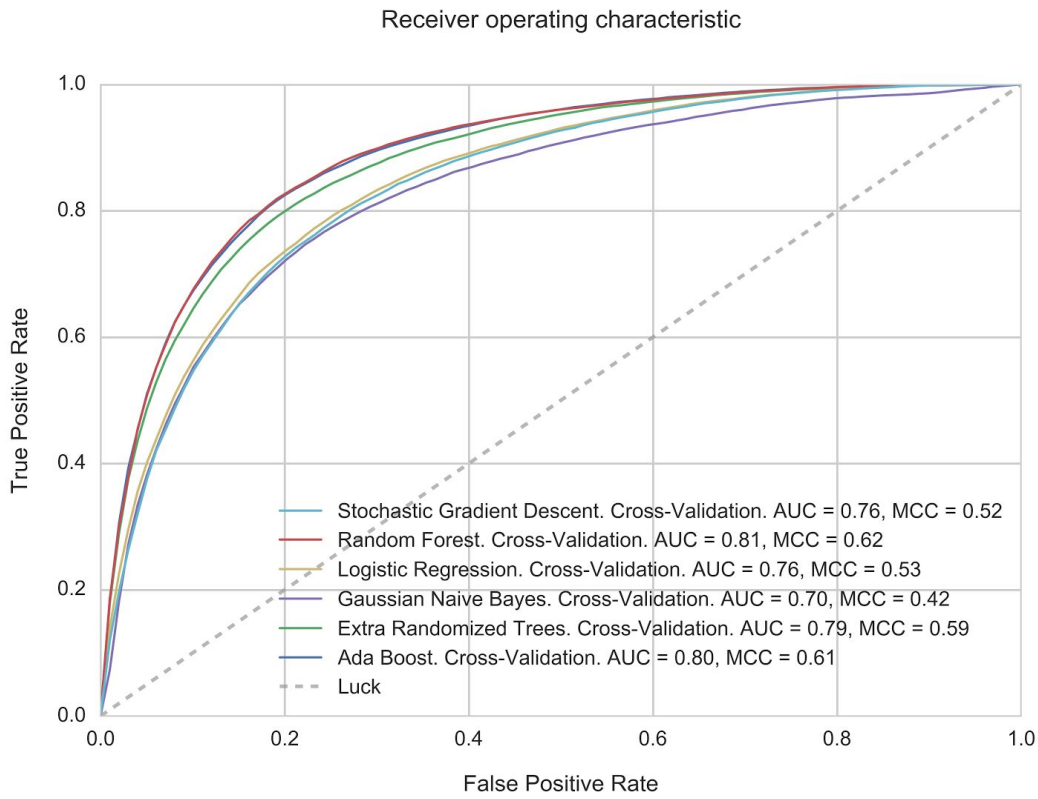

Figure S3. ROC Curves comparison of classifiers for PMut2017.
ROC curves of the classifier comparison in Table S4. Random Forest presents the best performance, closely followed by AdaBoost.

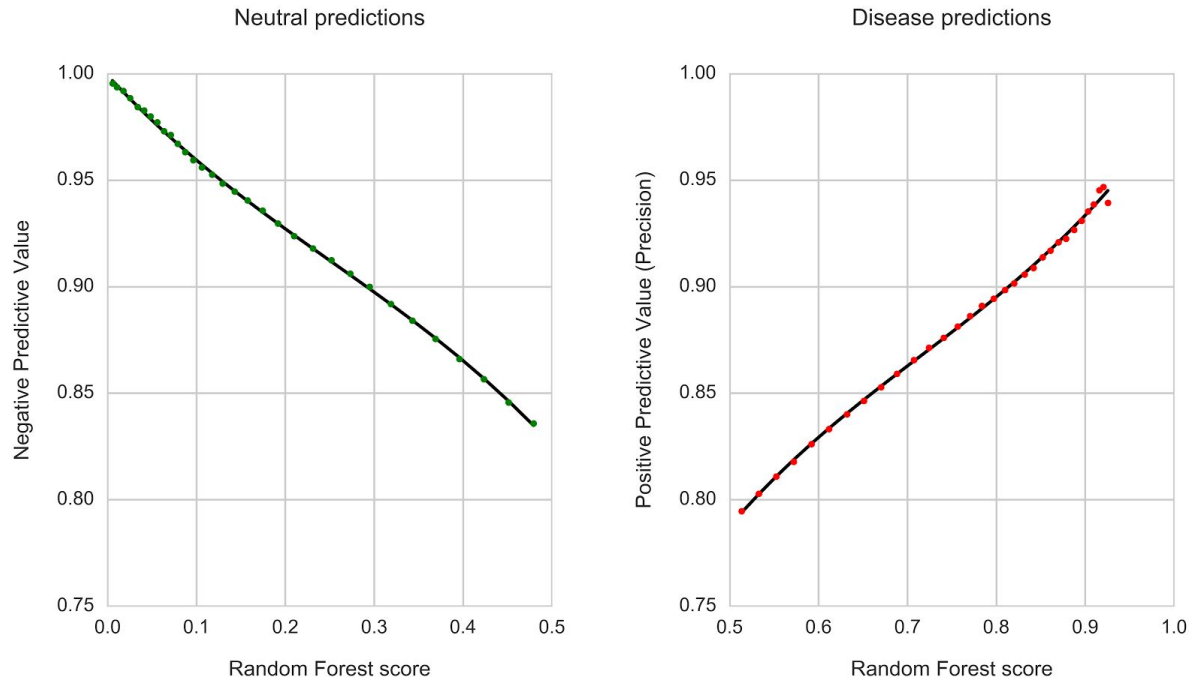

Figure S4. Reliability score regression for PMut2017.
We plot the Negative Predictive Value, $NPV = \frac{TN}{TN+FN}$ (left) and the Precision or Positive Predictive Value, $PPV = \frac{TP}{TP+FP}$ (right) for different thresholds of the Random Forest score; where TN, FN, TP, FP are the True Negatives, False Negatives, True Positives and False Positives in the predictions of a 10-fold cross-validation with 50% sequence identity exclusion.

We map the Random Forest score to NPV and PPV using an univariate spline regression. The NPV and the PPV give us the probability that a prediction is correct for the Neutral and Disease cases respectively.

REFERENCES

1. Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 10915-10919.
2. Dayhoff, M. O., R. M. Schwartz, and B. C. Orcutt. "A Model of Evolutionary Change in Proteins." *Atlas of protein sequence and structure*. Vol. 5. National Biomedical Research Foundation Silver Spring, MD, 1978. 345-352.
3. Miyata, T., Miyazawa, S. and Yasunaga, T. (1979) Two types of amino acid substitutions in protein evolution. *Journal of molecular evolution*, **12**, 219-236.
4. Chothia, C. (1975) Structural invariants in protein folding. *Nature*, **254**, 304-308.
5. Wimley, W.C. and White, S.H. (1996) Experimentally determined hydrophobicity scale for proteins at membrane interfaces. *Nature structural biology*, **3**, 842-848.
6. Fauchere, J.L., Charton, M., Kier, L.B., Verloop, A. and Pliska, V. (1988) Amino acid side chain parameters for correlation studies in biology and pharmacology. *International journal of peptide and protein research*, **32**, 269-278.
7. Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, **157**, 105-132.
8. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, **25**, 3389-3402
9. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B. and Wu, C.H. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics (Oxford, England)*, **31**, 926-932.
10. Lassmann, T. and Sonnhammer, E.L. (2005) Kalign--an accurate and fast multiple sequence alignment algorithm. *BMC bioinformatics*, **6**, 298.
11. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research*, **31**, 365-370.
12. Breiman, L., (2001) Random Forests. *Machine Learning*, **45**(1), 5-32
13. Freund, Y.; Schapire, R. (1995) A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*. **55,** 119-139
14. Geurts, P., Ernst, D.; Wehenkel, L. (2006) *Machine Learning* **63**, 3-42
15. Yip, Y.L., Famiglietti, M., Gos, A., Duek, P.D., David, F.P., Gateau, A. and Bairoch, A. (2008) Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. *Human mutation*, **29**, 361-366.